BEFORE THE OFFICE OF THE ASSISTANT SECRETARY FOR FAIR HOUSING AND EQUAL OPPORTUNITY, HUD

COMMENT REGARDING DOCKET NO. FR-6111-P-02

COMMENT OF CATHY O'NEIL

1. INTRODUCTION.

I submit this comment in response to the United States Department of Housing and Urban Development's (HUD's) proposed rule amending its disparate impact standard.[1]

HUD's call for "comments on the nature, propriety, and use of algorithmic models as related to" proposed defenses against disparate impact liability relates directly to my expertise.[2] I earned a Ph.D. in mathematics from Harvard University, completed my postdoctoral research in the MIT mathematics department, and worked as a professor at Barnard College, where I published a number of research papers in arithmetic algebraic geometry. I then transitioned to the private sector, working as a quantitative analyst for the hedge fund D.E. Shaw and then for RiskMetrics, a software company that assesses risk for hedge fund and bank holdings. In 2011, I left finance and started working as a data scientist, building models that predicted people's purchases and clicks. I am the author of *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* and a regular contributor to Bloomberg View where I comment on algorithmic justice. I also recently founded ORCAA, an algorithmic auditing company.

In *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*, the Supreme Court held that disparate impact claims are cognizable under the Fair Housing Act (FHA), reasoning that the FHA's language indicates an emphasis on the consequences of a practice rather than merely on the actor's intent.[3] It also established that plaintiffs bringing disparate impact claims face a robust causality requirement and must point to a specific practice or policy that directly results in the alleged disparity.[4] Furthermore, defendants must be given an opportunity to defend challenged practices by establishing that they serve valid interests.[5]

---

[1] HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. 42854 (proposed Aug. 19, 2019) (hereinafter "Proposed Rule").
[2] *Id.* at 42860.
[3] *See Texas Dep't of Hous. & Cmty. Affairs v. Inclusive Communities Project, Inc.*, 135 S. Ct. 2507, 2511 (2015).
[4] *See id.* at 2512.
[5] *See id.* at 2522.

The proposed rule purports to amend HUD's interpretation of the FHA's disparate impact standard in order to better reflect *Inclusive Communities*. The rule establishes five elements that plaintiffs must sufficiently plead in order to raise a disparate impact claim. Essentially, plaintiffs must show that the defendant is engaging in a policy or practice that is arbitrary, artificial, and unnecessary, and that there is a robust causal link between the challenged policy or practice and significant disparate impact on members of a protected class.[6]

In addition, the proposed rule provides three defenses for defendants who are accused of causing disparate impact via the use of algorithmic models.[7] HUD states that the proposed defenses are in response to comments it received "expressing concern that complicated, yet increasingly commonly used, algorithmic models to assess factors such as risk or creditworthiness, should be provided a safe harbor."[8]

The first defense protects a defendant who "[p]rovides the material factors that make up the inputs used in [a] challenged model and shows that these factors do not rely in any material part on factors that are substitutes or close proxies for protected classes under the Fair Housing Act and that the model is predictive of credit risk or other similar valid objective."[9]

The second defense indemnifies a defendant against disparate impact liability if the defendant "[s]hows that the challenged model is produced, maintained, or distributed by a recognized third party" whose tool is standard in the industry "and [that] the defendant is using the model as intended by the third party."[10]

Finally, the third defense allows a defendant to avoid liability if the defendant "shows that the model has been subjected to critical review and has been validated by . . . [a] third party that has analyzed the challenged model and found that the model . . . accurately predicts risk or other valid objectives, and that none of the factors used in the algorithm rely . . . on factors that are substitutes or close proxies for protected classes under the Fair Housing Act."[11]

I write to respond to HUD's request for comment on the three algorithmic defenses under the proposed rule. As written, the provisions—

---

[6] Proposed Rule at 42862.
[7] *See id.*
[8] *Id.* at 42859.
[9] *Id.* at 42862.
[10] *Id.*
[11] *Id.*

- wrongly assume that the exclusion of certain inputs is sufficient to ensure that a model cannot be the actual cause of disparate impact;

- contradict settled law holding that disparate impact liability is concerned with outcomes; and

- incentivize against testing for algorithmic bias.

For these reasons, I discourage HUD from adopting the proposed algorithmic defenses.

2. MODELS MAY CAUSE DISPARATE IMPACT EVEN IF PROXIES FOR PROTECTED CLASSES ARE EXCLUDED.

Under the first and third defenses proposed in §100.500(c)(2), a defendant may avoid disparate impact liability by demonstrating (or, in the case of the third defense, having a neutral third party attest) that none of the material input factors in a challenged algorithmic model used by the defendant represent substitutes or close proxies for protected classes under the FHA and that the model is predictive of risk or serves some other valid objective.[12] HUD reasons that, if a defendant is able to show that the individual factors used in a model are not substitutes or close proxies for characteristics of protected classes, then the algorithm cannot be the actual cause of the disparate impact alleged by the plaintiff.[13]

HUD's reasoning on this point is incorrect. There may be a robust causal link, as required in *Inclusive Communities,* between a challenged model and alleged disparate impact even where the factors making up the inputs used in the model do not rise to the level of substitutes or close proxies. Policing individual variables has limited efficacy in addressing the issues raised by the use of increasingly sophisticated algorithms in decision-making. Complex models will systematically pick up on characteristics of protected classes, even if certain inputs are excluded. Furthermore, human choices regarding how target variables are defined, which features are deemed important to include in a model, and how data is collected may result in disparate impact that cannot be addressed by merely excluding inputs. Consequently, defendants should not automatically be able to avoid liability under the proposed defenses.

2.1 Variables may be correlated with protected characteristics even if they do not constitute traditional proxies or substitutes, and models may still pick up on protected

[12] *Id.*
[13] *See id.* at 42859.

characteristics even when correlated inputs are intentionally excluded.

As a threshold matter, it is difficult to define the criteria under which variables are too closely related to protected characteristics to be included as inputs to a model. The proposed rule does not address the question of how heavily other input variables must correlate with those protected traits in order to place a model outside the protection of paragraph (c)(2). It merely proscribes "substitutes" and "close proxies," neither of which are defined or quantified.[14] Variables may be correlated with protected characteristics[15] while not rising to the level of substitutes or close proxies (assuming those terms are interpreted to mean "traditional" proxies with a very high correlation to protected characteristics).[16]

For example, under the proposed rule, landlords may be able to avoid liability if the algorithms they use to screen tenants do not take into account obvious proxies for race, such as zip code, even if, by mining applicants' streaming data or social media activity, the algorithms analyze factors that are less intuitively or reliably correlated with race, such as musical genre preference.[17] An algorithm that picks up on a relationship between preference for hip-hop music and frequency of noise complaints, consequently using music preference as a factor in its predictions, is probably causing disparate impact based on race.[18] Eliminating substitutes and close proxies for protected characteristics from such algorithmic models is not sufficient to ensure that the models cannot still pick up on race and thus have discriminatory effects. This is especially true in the context of complex artificial intelligence algorithms designed to examine the "interaction *between* features to find unexpected patterns in the data."[19]

Even if the threshold for substitutes and close proxies could be quantified, the fact that no *individual* inputs to a model are significantly correlated with protected classes does not mean that a *combination* of many variables cannot *jointly* code for protected characteristics. Thus, focusing on the

---

[14] Proposed Rule at 42862.
[15] Some scholars argue that it is a "statistical reality" that virtually all algorithmic inputs are at least somewhat correlated with race. *See* Crystal Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework* 6 (October 1, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3462379.
[16] Note that, in describing the first defense, HUD suggests that a plaintiff may show "that a factor used in the model is correlated with a protected class despite the defendant's assertion." *Id.* at 42859. However, the text of §100.500 itself refers only to "substitutes" or "close proxies," not mere correlation. *Id.* at 42862.
[17] *See* Andrew D. Selbst, *A New HUD Rule Would Effectively Encourage Discrimination by Algorithm*, Slate (Aug. 19, 2019), https://slate.com/technology/2019/08/hud-disparate-impact-discrimination-algorithm.html.
[18] *See* i*d.*
[19] *Id.* (emphasis added).

exclusion of individual inputs is a misguided approach. The more complex the algorithm, the more pronounced this problem becomes. For example, the class of algorithms categorized as artificial intelligence (AI) "use training data to discover on their own what characteristics can be used to predict the target variable," a process that "results in AIs inevitably 'seeking out' proxies for directly predictive characteristics when direct data on these characteristics is not made available to the AI due to legal prohibitions."[20] Because AIs are designed to discover patterns and relationships in data that can help predict factors such as risk or creditworthiness, forbidding the inclusion of any number of inputs to an algorithmic model is unlikely to significantly reduce discriminatory potential.[21] If the AIs are deprived of the predictive power of forbidden variables, they will simply reconstruct these variables using other available data. Furthermore, with the increased availability of non-traditional data (for example, social media profiles), there is also an increased likelihood that an algorithm will be able to recover and infer protected characteristics.

The effectiveness of some algorithms at reconstructing protected traits despite the exclusion of certain inputs is demonstrated in a study conducted by scholars Talia B. Gillis and Jann Spiess.[22] As part of their study, Gillis and Spiess observed that predicted mortgage default distributions were different for samples of whites, blacks, and Hispanics even when race was excluded as an input from the algorithm producing the predictions. Furthermore, this same disparity persisted (though it was smaller) even when the ten variables that correlate most strongly with race

---

[20] Anya Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, Iowa. L. Rev. 1, 8 (forthcoming 2020).

[21] Because excluding protected traits and their proxies does not prevent those traits from entering into a model, some scholars argue that models should include protected class variables and subsequently adjust for them for optimal fairness. *See* Yang & Dobbie, *supra*, at 33 ("Unlike the excluding-inputs algorithm . . . the colorblinding-inputs algorithm does not exclude race and race-correlates in the estimation step. In fact, it uses all inputs to estimate predictive relationships, in contrast to the current approach of using ad hoc human judgment to decide which correlated inputs should be included or excluded."); *See also* Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 Wm. & Mary L. Rev. 857, 904 (2017)(" . . . a blanket prohibition on the explicit use of race or other prohibited characteristics does not avoid, and may even worsen, the discriminatory impact of relying on a data model."); Prince & Schwarcz, *supra*, at 63-64 ("Counterintuitively, the first step . . . is for the statistical model under consideration to be re-estimated in a way that explicitly includes data on legally prohibited characteristics. For a model produced by an AI, accomplishing this requires including in the training data information on legally prohibited characteristics, such as the race or health status of individuals in the training population. This first step is necessary because it removes from all of the legally-permitted variables any predictive power that derives from those variables' capacity to proxy for a prohibited characteristic.").

[22] *See* Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. Chi. L. Rev 459, 468-470.

were also excluded.[23] Gillis and Spiess noted that, "in big data, even excluding those variables that individually relate most to the 'forbidden input' does not necessarily significantly affect how much pricing outputs vary with, say, race."[24]

As the Gillis and Spiess study suggests, a legal doctrine requiring plaintiffs to identify the specific inputs causing disparate impact is inadequate in a world of big data and artificial intelligence. It is virtually impossible to "determine whether an AI is proxying for a protected trait simply by scrutinizing the data on which it ultimately relies," when "the proxies available to AIs may consist of numerous interacting pieces of data, whose significance as a proxy may be completely unintuitive."[25] The fact that models produced by such algorithms may be so complicated and sophisticated as to be inscrutable *even to those who use them*[26] weighs in favor of a more holistic and outcome-oriented (as opposed to input-oriented) causality standard in this particular subset of disparate impact claims.

2.2 Human decisions regarding target variable definition, feature selection, and data collection may cause disparate impact, and excluding inputs will do little to alleviate such discrimination.

Even with simpler algorithms, excluding substitutes or close proxies from a model is unlikely to ensure that a model will not result in disparate impact. This is because all algorithms reflect human goals and ideology, "from the data we choose to collect to the questions we ask . . . models are opinions embedded in mathematics."[27] The fact that humans impose ideologies on models has far greater discriminatory potential than simply the possibility that protected characteristics or proxies for protected characteristics might be used as inputs.

For one, defining the target variable of interest often involves discretion on the part of developers. Developers must "translate some amorphous problem," such as how to hire good employees or how to extend credit, into something that "can be expressed in more formal terms that computers can parse."[28] An algorithm cannot predict, for example, who is likely to be a good employee without first being provided with a definition of "good" that "correspond[s] to measurable outcomes." While "good" is not directly

---

[23] *Id.* at 469-470.
[24] *Id.* at 470.
[25] Prince & Schwarcz, *supra*, at 52-53.
[26] *See id.* at 7 n.14.
[27] Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy 29 (2016).
[28] *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671, 678 (2016).

measurable, "relatively higher sales, shorter production time, or longer tenure" *can* be measured, and each represents only one possible definition of what it means to be a "good" employee.[29] Similarly, creditworthiness is not directly measurable and must be subjectively defined in measurable ways. Because there is an element of arbitrariness to target variable definition, the process might very well reflect problematic assumptions that result in disparate impact on members of protected classes. For example, Solon Barocas and Andrew Selbst write that:

> Hiring decisions made on the basis of predicted tenure are much more likely to have a disparate impact on certain protected classes than hiring decisions that turn on some estimate of worker productivity. If the turnover rate happens to be systematically higher among members of certain protected classes, hiring decisions based on predicted length of employment will result in fewer job opportunities for members of these groups, even if they would have performed as well as or better than the other applicants the company chooses to hire.[30]

Furthermore, because there is no way for employers to learn how many good employees they have missed out on by choosing one definition of "good" over another, they are unlikely to have any incentive to revise their definitions or rethink their assumptions. Likewise, defenses centered on the exclusion of individual inputs do not add any incentives for housing authorities to question any of their assumptions that may be resulting in disparate impact.

Take another example of problematic target variable definition involving algorithmic risk assessment tools in the criminal justice system:

> Statistical validation of recidivism prediction in particular suffers from a fundamental problem: . . . since the target for prediction (having actually committed a crime) is unavailable, it is tempting to change the goal of the tool to predicting arrest, rather than crime . . . One problem with using such imperfect proxies is that different demographic groups are stopped, searched, arrested, charged, and are wrongfully convicted at very different rates in the current US criminal justice system. Further, different types of crimes are reported and recorded at different rates, and the rate

[29] *Id.* at 679.
[30] *Id.* at 680.

> of reporting may depend on the demographics of the perpetrator and victim.[31]

When there are challenges in directly measuring some outcome of interest, those who develop algorithmic tools must define a target variable that is measurable. These choices may have the effect of disadvantaging protected classes, and sanitizing algorithm inputs would not alleviate the issue.

An algorithm may also cause disparate impact if its design does not factor in variables that would more accurately predict whether a member of a protected class possesses some desired characteristic, such as creditworthiness. For example, for members of groups who have historically been excluded from the traditional credit system, factors such as timeliness of rent payments, phone bills, and utilities, which are not included in credit scoring parameters, may more accurately reflect creditworthiness than length of credit history, which makes up a large percentage of a traditional credit score.[32] Issues arise when members of protected classes "are subject to systematically less accurate classifications or predictions because the details necessary to achieve equally accurate determinations reside at a level of granularity and coverage that the selected features fail to achieve." [33] Decisions regarding which particular factors to include are inevitable, so models will invariably have blind spots. But blind spots that systematically disadvantage members of protected classes should not be tolerated, and the mere exclusion of certain inputs from a model will not remove them.

Finally, a model may result in disparate impact if the data on which it bases its predictions is incomplete, inaccurate, or unrepresentative. Barocas and Selbst note that "the quality and representativeness of records might vary in ways that correlate with class membership (e.g., institutions might maintain systematically less accurate, precise, timely, and complete records for certain classes of people)."[34]

In summary, there are myriad ways in which models may result in disparate impact even when certain factors are excluded as inputs. HUD is thus

---

[31] Partnership on AI, *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System*, https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/(last visited Oct. 18, 2019).
[32] Lauren deLisa Coleman, *Inside the Alarming Way the Underbelly of Algorithms is Strangling the American Dream*, Forbes (Aug. 27, 2019), https://www.forbes.com/sites/laurencoleman/2019/08/27/inside-the-alarming-way-the-underbelly-of-algorithms-is-strangling-the-american-dream/#7e4169e06d2f.
[33] Barocas & Selbst, *supra*, at 688.
[34] *Id.* at 684.

mistaken in assuming that the input-focused elements of its proposed defenses are sufficient to prove that a model is not the actual cause of alleged disparate impact.

3. EXAMINING INPUTS OVER OUTCOMES IS INCONSISTENT WITH DISPARATE IMPACT LAW.

Defenses that focus on the exclusion of proxies for protected classes also run counter to disparate impact doctrine. Courts impose disparate impact liability when a defendant's policy or practice results in an unjustified discriminatory outcome.[35] The fact that a defendant did not directly consider protected characteristics when they made their decision does not excuse liability in disparate impact analysis. Therefore, defenses against disparate impact liability that focus on whether an algorithmic model uses racial proxies are inconsistent with the law because they police inputs rather than outcomes.

Under Supreme Court precedent, disparate impact liability arises even when protected characteristics do not factor into decision-making. The Court first endorsed disparate impact claims in *Griggs v. Duke Power*.[36] There, an employer would not hire manual laborers without high school diplomas even though high school education was unrelated to job performance. At the time, blacks were less likely than whites to possess a high school diploma, so the hiring policy disproportionately screened out black applicants. The Court concluded that this disparate impact was unacceptable under Title VII because it created "artificial, arbitrary, and unnecessary barriers to employment" for blacks.[37] Although the diploma policy was race-neutral and therefore "fair in form," it nonetheless violated Title VII because it "discriminat[ed] in operation."[38] Applying the logic of *Griggs*, the Supreme Court held in *Inclusive Communities* that the FHA also prohibits disparate impact discrimination.[39] The opinion once again affirmed the principle that disparate impact liability is "results-oriented."[40] In other words, the FHA emphasizes consequences over procedural fairness.

Although the proposed rule purports to bring HUD's disparate impact standard in line with *Inclusive Communities*, HUD's focus on algorithmic inputs does the opposite.[41] Under the proposed rule, defendants are not liable for an algorithm's discriminatory effects so long as defendants show

---

[35] *See Inclusive Communities,* 135 S. Ct. at 2512.
[36] *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971).
[37] *Id.* at 431.
[38] *Griggs,* 401 U.S. at 431.
[39] *Inclusive Communities*, 135 S. Ct. at 2512.
[40] *Id.* at 2511.
[41] *See* Proposed Rule at 42854.

that the algorithm does not consider proxies for protected classes as inputs.[42] Given *Inclusive Communities*' focus on outcomes, this approach is misguided. Algorithms that do not use protected characteristics as inputs but nonetheless produce unjustifiable discriminatory results are no different than the diploma policy in *Griggs*: "fair in form, but discriminatory in operation."[43]

HUD may receive comments mistakenly suggesting that disparate impact liability exists only to combat veiled or subconscious direct discrimination. Under this "evidentiary dragnet" view of disparate impact liability, disparate impact law focuses on discriminatory effect merely to catch insidious forms of prejudice that would be too difficult to prove.[44] But, in the context of the FHA, the purpose of disparate impact litigation is broader than simply rooting out subconscious direct discrimination.

As the Court explains in *Inclusive Communities*, Congress passed the FHA not only to combat housing discrimination but also to promote residential desegregation.[45] In February 1968, the National Advisory Commission on Civil Disorders released a report that identified unequal housing, residential segregation, and economic inequality as animating forces behind race riots.[46] To reduce social unrest, the Commission "recommended enactment of 'a comprehensive and enforceable open-occupancy law making it an offense to discriminate in the sale or rental of any housing . . . on the basis of race, creed, color, or national origin.'"[47] Less than two months later, Dr. Martin Luther King was assassinated, and riots ensued. "Congress responded by . . . passing the FHA"[48] just days after Dr. King's death, making it illegal to deny housing on account of "race, color, religion, or national origin."[49]

The FHA's origin story demonstrates that its aims are too ambitious for the "evidentiary dragnet" view of disparate impact to hold. Congress passed the FHA to ease racial tension by moving America towards a more equal integrated society. In this light, disparate impact liability exists as an aggressive tool "to dismantle racial hierarchies regardless of whether

---

[42] *See id.* at 42862
[43] *Griggs,* 401 U.S. at 431.
[44] *See* Richard A. Primus, Equal Protection and Disparate Impact: Round Three, 117 Harv. L. Rev. 493, 520 (2003).
[45] *See Inclusive Communities*, 135 S. Ct. at 2516.
[46] *See* Report of the National Advisory Commission on Civil Disorders 91 (1968) (Kerner Commission Report).
[47] *See Inclusive Communities,* 135 S. Ct. at 2516.
[48] *Id.* at 2516.
[49] Civil Rights Act of 1968, § 804, 82 Stat. 83. Other protected classes were added under Fair Housing Amendments Act of 1988, 102 Stat. 1619.

anything like intentional discrimination is present."[50] Disparate impact doctrine orients liability around unjustified discriminatory effects because it was designed to not only stop discrimination but to promote active integration by forcing the housing industry to consider the discriminatory effects of its policies. To stay true to the FHA's results-oriented framework, HUD should abandon its attempt to excuse disparate impact liability based on algorithmic inputs.

4. IMMUNIZING DEFENDANTS WHO USE INDUSTRY-STANDARD TOOLS DISINCENTIVIZES TESTING FOR ALGORITHMIC BIAS.

HUD should also abandon the second algorithmic defense in paragraph (c)(2) of §100.500 because it incentivizes against testing for algorithmic bias. The defense immunizes defendants from disparate impact liability related to the use of an algorithm if the algorithm in question was developed by a third party whose tool is standard in the industry.[51] In effect, the provision would push both housing providers and algorithm vendors to ignore disparate impact risk.[52]

For lenders, landlords, and other housing industry players, the proposed rule creates what economists refer to as moral hazard: a situation where a party is willing to increase exposure to risk because another party bears the cost.[53] Without the threat of disparate impact liability, any housing industry player using a third-party algorithm will be willing to take the risk that its purchased algorithm discriminates because the HUD policy shifts the burden of disparate impact liability onto the algorithm vendor. Banks and landlords have no reason to consider whether the algorithms they purchase disparately impact members of protected classes if they face no legal risk from using biased models. Therefore, testing third-party algorithms for disparate impact on protected groups becomes an unnecessary expense under the proposal.

The proposed rule also disincentivizes algorithm vendors from testing for bias. Without the threat of disparate impact liability, algorithm buyers will not demand unbiased algorithms. And in turn, vendors will have no market incentive to create unbiased algorithms. Although the threat vendors face from disparate impact liability should incentivize them to develop fairer algorithms, this depends on the nature of the vendors themselves. When defendants cannot afford to pay judgments, plaintiffs will not bring suit. In

---

[50] Primus, *supra*, at 518.
[51] Proposed Rule at 42862.
[52] Kriston Capps, *How HUD Could Dismantle a Pillar of Civil Rights Law*, CityLab (Aug. 16, 2019), https://www.citylab.com/equity/2019/08/fair-housing-act-hud-disparate-impact-discrimination-lenders/595972/.
[53] Mark Thoma, *Explainer: What is "moral hazard"?*, CBS News (Nov. 22, 2013), https://www.cbsnews.com/news/explainer-moral-hazard/.

the context of the FHA, plaintiffs will not bring disparate impact claims when algorithm vendors are small businesses without much capital. Liability should lie with entities that purchase algorithms because their market power can force even small, judgement-proof algorithm vendors to develop fairer models.

Neither of HUD's justifications for the third-party defense can withstand scrutiny. First, HUD suggests that providing immunity to defendants who use industry standard third-party algorithms is fair because "the defendant may not have access to the reasons these factors are used or may not even have access to the factors themselves, and, therefore, may not be able to defend the model itself."[54] But ignorance is not an excuse from liability under disparate impact law. If lenders who use third-party algorithms are worried about defending against disparate impact liability, then they should discontinue using opaque algorithms and demand transparency.

Second, HUD claims that its proposed provision would be more efficient. Successful suits under the proposal would presumably remove biased algorithms from the market entirely because the vendors themselves would be liable rather than any individual entity using a model. This justification fails because algorithm vendors already face liability for disparate impact claims under current HUD regulations. As explained in a recent case, "HUD regulation [creating] liability for a person's 'own conduct that results in a discriminatory housing practice'" already imposes liability on vendors selling biased algorithms.[55] Additionally, even if plaintiffs sue algorithm users rather than vendors, housing industry players will voluntarily discontinue using a biased algorithm once a successful suit demonstrates that the model is a liability.

5. CONCLUSION.

Ultimately, HUD's proposed disparate impact rule betrays fundamental misunderstandings about the nature of algorithmic bias, disparate impact law, and economic incentives. Excluding proxies for protected class status does little to solve the problem of algorithmic bias in a world where HUD fails to define what counts as a proxy and where modern tools re-encode protected characteristics despite developers' best efforts. *Inclusive Communities*—the very case that spurred HUD to rewrite its disparate impact regulations—reaffirms that discriminatory outcome is the touchstone of disparate impact analysis. Algorithmic defenses that focus on inputs run counter to that settled precedent. Finally, indemnifying those

[54] *See* Proposed Rule at 42859.
[55] *Connecticut Fair Hous. Ctr. v. Corelogic Rental Prop. Sols., LLC,* 369 F. Supp. 3d 362, 372 (D. Conn. 2019) (citing 24 C.F.R. § 100.7(a)(iii)).

who use third-party algorithms from liability significantly reduces incentives for algorithm users and vendors to test their tools for bias.

Respectfully submitted,[56]

Christopher Bavitz
Managing Director, Cyberlaw Clinic
Mason Kortz
Clinical Instructor, Cyberlaw Clinic
Tea Skela
Clinical Law Student, Cyberlaw Clinic (F19)
James Holloway
Clinical Law Student, Cyberlaw Clinic (F19)
Harvard Law School
1585 Massachusetts Avenue, Suite WCC 5018
Cambridge, MA 02138
Tel: 617-394-9125
Email: cbavitz@law.harvard.edu

On behalf of Cathy O'Neil

[56] Thanks to Berkman Klein Center for Internet & Society Project Coordinator Adam Nagy for his valuable contributions to this comment.